

RECURSIVE BIAS ESTIMATION AND L_2 BOOSTING

BY PIERRE-ANDRÉ CORNILLON, NICOLAS HENGARTNER AND ERIC
MATZNER-LØBER

*Montpellier SupAgro, University Rennes 2 and Los Alamos National
Laboratory*

This paper presents a general iterative bias correction procedure for regression smoothers. This bias reduction schema is shown to correspond operationally to the L_2 Boosting algorithm and provides a new statistical interpretation for L_2 Boosting. We analyze the behavior of the Boosting algorithm applied to common smoothers S which we show depend on the spectrum of $I - S$. We present examples of common smoother for which Boosting generates a divergent sequence. The statistical interpretation suggest combining algorithm with an appropriate stopping rule for the iterative procedure. Finally we illustrate the practical finite sample performances of the iterative smoother via a simulation study. simulations.

1. Introduction. Regression is a fundamental data analysis tool for uncovering functional relationships between pairs of observations $(X_i, Y_i), i = 1, \dots, n$. The traditional approach specifies a parametric family of regression functions to describe the conditional expectation of the dependent variable Y given the independent variables $X \in \mathbb{R}^p$, and estimates the free parameters by minimizing the squared error between the predicted values and the data. An alternative approach is to assume that the regression function varies smoothly in the independent variable x and estimate locally the conditional expectation of Y given X . This results in nonparametric regression estimators (e.g. Fan and Gijbels [13], Hastie and Tibshirani [19], Simonoff [34]). The vector of predicted values \hat{Y}_i at the observed covariates X_i from a nonparametric regression is called a regression smoother, or simply a smoother, because the predicted values \hat{Y}_i are less variable than the original observations Y_i .

Over the past thirty years, numerous smoothers have been proposed: running-mean smoother, running-line smoother, bin smoother, kernel based smoother (Nadaraya [29], Watson [38]), spline regression smoother, smoothing splines smoother (Wahba [37], Whittaker [39]), locally weighted running-line smoother (Cleveland [6]), just to mention a few. We refer to Buja et al.

AMS 2000 subject classifications: 62G08

Keywords and phrases: nonparametric regression, smoother, kernel, nearest neighbor, smoothing splines, stopping rules

[5], Eubank [12], Fan and Gijbels [13], Hastie and Tibshirani [19] for more in depth treatments of regression smoothers.

An important property of smoothers is that they do not require a rigid (parametric) specification of the regression function. That is, we model the pairs (X_i, Y_i) as

$$(1.1) \quad Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where $m(\cdot)$ is an unknown smooth function. The disturbances ε_i are independent mean zero and variance σ^2 random variables that are independent of the covariates X_i , $i = 1, \dots, n$. To help our discussion on smoothers, we rewrite Equation (1.1) compactly in vector form by setting $Y = (Y_1, \dots, Y_n)^t$, $m = (m(X_1), \dots, m(X_n))^t$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^t$, to get

$$(1.2) \quad Y = m + \varepsilon.$$

Finally we write $\hat{m} = \hat{Y} = (\hat{Y}_1, \dots, \hat{Y}_n)^t$, the vector of fitted values from the regression smoother at the observations. Operationally, linear smoothers can be written as

$$\hat{m} = S_\lambda Y,$$

where S_λ is a $n \times n$ smoothing matrix. While in general the smoothing matrix will be not be a projection, it is usually a contraction (Buja et al. [5]). That is, $\|S_\lambda Y\| \leq \|Y\|$.

Smoothing matrices S_λ typically depend on a tuning parameter, which denoted by λ , that governs the tradeoff between the smoothness of the estimate and the goodness-of-fit of the smoother to the data. We parameterize the smoothing matrix such that large values of λ will produce very smooth curves while small λ will produce a more wiggly curve that wants to interpolate the data. The parameter λ is the bandwidth for kernel smoother, the span size for running-mean smoother, bin smoother, and the penalty factor λ for spline smoother.

Much has been written on how to select an appropriate smoothing parameter, see for example (Simonoff [34]). Ideally, we want to choose the smoothing parameter λ to minimize the expected squared prediction error. But without explicit knowledge of the underlying regression function, the prediction error can not be computed. Instead, one minimizes estimates of

the prediction error using Stein Unbiased Risk Estimate or Cross-Validation (Li [26]).

This paper takes a different approach. Instead of selecting the tuning parameter λ , we fix it to some reasonably large value, in a way that ensures that the resulting smoothers *oversmooths* the data, that is, the resulting smoother will have a relatively small variance but a substantial bias. Observe that the conditional expectation of the $-R = -(Y - \hat{Y})$ given X is the bias of the smoother. This provides us with the opportunity of estimating the bias by smoothing the residuals R , thereby enabling us to bias correct the initial smoother by subtracting from it the estimated bias. The idea of estimating the bias from residuals to correct a pilot estimator of a regression function goes back to the concept of *twicing* introduced by (Tukey [35]) to estimate bias from model misspecification in multivariate regression. Obviously, one can iteratively repeat the bias correction step until the increase to the variance from the bias correction outweighs the magnitude of the reduction in bias, leading to an iterative bias correction.

Another iterative function estimation method, seemingly unrelated to bias reduction, is Boosting. Boosting was introduced as a machine learning algorithm for combining multiple weak learners by averaging their weighted predictions (Freund [15], Schapire [31]). The good performance of the Boosting algorithm on a variety of datasets stimulated statisticians to understand it from a statistical point of view. In his seminal paper, Breiman [2] shows how Boosting can be interpreted as a gradient descent method. This view of Boosting was reinforced by Friedman [16]. Adaboost, a popular variant of the Boosting algorithm, can be understood as a method for fitting an additive model (Friedman et al. [17]) and recently Efron et al. [11] made a connection between L_2 Boosting and Lasso for linear models.

But connections between iterative bias reduction and Boosting can be made. In the context of nonparametric density estimation, Di Marzio and Taylor [8] have shown that one iteration of the Boosting algorithm reduced the bias of the initial estimator in a manner similar to the multiplicative bias reduction methods (Hengartner and Matzner-Løber [20], Hjort and Glad [22], Jones et al. [25]). In the follow-up paper (Di Marzio and Taylor [9]), they extend their results to the nonparametric regression setting and show that one step of the Boosting algorithm applied to an oversmooth effects a bias reduction. As expected, the decrease in the bias comes at the cost of an increase in the variance of the corrected smoother.

In Section 2, we show that in the context of regression, such iterative bias reduction schemes obtained by correcting an estimator by smoothers of

the residuals correspond operationally to the L_2 Boosting algorithm. This provides a novel statistical interpretation of L_2 Boosting. This new interpretation helps explain why, as the number of iteration increases, the estimator eventually deteriorates. Indeed, by iteratively reducing the bias, one eventually adds more variability than one reduces the bias.

In Section 3, we discuss the behavior of the L_2 Boosting of many commonly used smoothers: smoothing splines, Nadaraya-Watson kernel and K -nearest neighbor smoothers. Unlike the good behavior of the L_2 boosted smoothing splines discussed in Buhlmann and Yu [4], we show that Boosting K -nearest neighbor smoothers and kernel smoothers that are not positive definite produces a sequence of smoothers that behave erratically after a small number of iteration, and eventually diverge. The reason for the failure of the L_2 Boosting algorithm, when applied to these smoothers, is that the bias is overestimated. As a result, the Boosting algorithm over-corrects the bias and produces a divergent smoother sequence. Section 4 discusses modifications to the original smoother to ensure good behavior of the sequence of boosted smoothers.

To control both the over-fitting and over-correction problems, one needs to stop the L_2 Boosting algorithm in a timely manner. Our interpretation of the L_2 Boosting as an iterative bias correction scheme leads us to propose in Section 5 several data driven stopping rules: Akaike Information Criteria (AIC), a modified AIC, Generalized Cross Validation (GCV), one and L -fold Cross Validation, and estimated prediction error estimation using data splitting. Using either the asymptotic results of Li [27] or the finite sample oracle inequality of Hengartner et al. [21], we see that stopped boosted smoother has desirable statistical properties. We use either of these theorems to conclude that the desirable properties of the boosted smoother does not depend on the initial pilot smoother, provided that the pilot oversmooths the data. This conclusion is reaffirmed from the simulation study we present in Section 6. To implement these data driven stopping rules, we need to calculate predictions of the smoother for any desired value of the covariates, and not only at the observations. We show in Section 5 how to extend linear smoothers to give predictions at any desired point.

The simulations in Section 6 show that when we combine a GCV based stopping rule to the L_2 Boosting algorithm seems to work well. It stops early when the Boosting algorithm misbehaves, and otherwise takes advantage of the bias reduction. Our simulation compares optimum smoothers and optimum iterative bias corrected smoothers (using generalized cross validation) for general smoothers without knowledge of the underlying regression function. We conclude that the optimal iterative bias corrected smoother

outperforms the optimal smoother.

Finally, the proofs are gathered in the Appendix.

2. Recursive bias estimation. In this section, we define a class of iteratively bias corrected linear smoothers and highlight some of their properties.

2.1. Bias Corrected Linear Smoothers. For ease of exposition, we shall consider the univariate nonparametric regression model in vector form (1.2) from Section 1

$$Y = m + \varepsilon,$$

where the errors ε are independent, have mean zero and constant variance σ^2 , and are independent of the covariates $X = (X_1, \dots, X_n)$, $X_j \in \mathbb{R}$. Extensions to multivariate smoothers are strait forward and we refer to Buja et al. [5] for example.

Linear smoothers can be written as

$$(2.1) \quad \hat{m}_1 = SY,$$

where S is an $n \times n$ smoothing matrix. Typical smoothing matrices are contractions, so that $\|SY\| \leq \|Y\|$, and as a result the associated smoother SY is called a shrinkage smoother (see for example Buja et al. [5]). Let I be the $n \times n$ identity matrix.

The linear smoother (2.1) has bias

$$(2.2) \quad B(\hat{m}_1) = \mathbb{E}[\hat{m}_1|X] - m = (S - I)m$$

and variance

$$V(\hat{m}_1|X) = SS'\sigma^2,$$

respectively.

A natural question is “how can one estimate the bias?” To answer this question, observe that the residuals $R_1 = Y - \hat{m}_1 = (I - S)Y$ have expected value $\mathbb{E}[R_1|X] = m - \mathbb{E}[\hat{m}_1|X] = (I - S)m = -B(\hat{m}_1)$. This suggests estimating the bias by smoothing the negative residuals

$$(2.3) \quad \hat{b}_1 := -SR_1 = -S(I - S)Y.$$

This bias estimator is zero whenever the smoothing matrix S is a projection, as is the case for linear regression, bin smoothers and regression splines.

However, since most common smoothers are not projections, we have an opportunity to extract further signal from the residual and possibly improve upon the initial estimator.

Note that a smoothing matrix other than S can be used to estimate the bias in (2.3), but as we shall see, in many examples, using S works very well, and leads to an attractive interpretation of Equation (2.3). Indeed, since the matrices S and $I - S$ commute, we can express the estimated bias as

$$\hat{b}_1 = -S(I - S)Y = -(I - S)SY = (S - I)\hat{m}_1.$$

We recognize the latter as the right-hand side of (2.2) with the smoother \hat{m}_1 replacing the unknown vector m . This says that \hat{b}_1 is a plug-in estimate for the bias $B(\hat{m}_1)$.

Subtracting the estimated bias from the initial smoother \hat{m}_1 produces the *twicing* estimator

$$\begin{aligned}\hat{m}_2 &= \hat{m}_1 - \hat{b}_1 \\ &= (S + S(I - S))Y \\ &= (I - (I - S)^2)Y.\end{aligned}$$

Since the twiced smoother \hat{m}_2 is also a linear smoother, one can repeat the above discussion with \hat{m}_2 replacing \hat{m}_1 , producing a *thriced* linear smoother. We can iterate the bias correction step to recursively define a family of bias corrected smoothers. Starting with $\hat{m}_1 = SY$, construct recursively for $k = 2, 3, \dots$, the sequences of residuals, estimated bias and bias corrected smoothers

$$\begin{aligned}R_{k-1} &= (I - S)^{k-1}Y \\ \hat{b}_k &= -SR_{k-1} = -(I - S)^{k-1}SY \\ (2.4) \quad \hat{m}_k &= \hat{m}_{k-1} - \hat{b}_k = \hat{m}_{k-1} + SR_{k-1}.\end{aligned}$$

We show in the next theorem that the iteratively bias corrected smoother \hat{m}_k defined by Equation 2.4 has a nice representation in terms of the original smoothing matrix S .

THEOREM 2.1. *The k^{th} iterated bias corrected linear smoother \hat{m}_k (2.4) can be explicitly written as*

$$\begin{aligned}\hat{m}_k &= S[I + (I - S) + (I - S)^2 + \dots + (I - S)^{k-1}]Y \\ (2.5) \quad &= [I - (I - S)^k]Y = S_k Y.\end{aligned}$$

Example with a Gaussian kernel smoother Throughout the next two sections, we shall use the following example to illustrate the behavior of the Boosting algorithms applied to various common smoothers. Take the design points to be 50 independently drawn points from an uniform distribution on the unit interval $[0, 1]$. The true regression function is $m(x) = \sin(5\pi x)$, the solid line in the Figure 1, and the disturbances are mean zero Gaussians with variance producing a signal to noise ratio of five.

In the next figure, the initial smoother is a kernel one, with a bandwidth equals to 0.2 and a Gaussian kernel. This pilot smoother heavily oversmooths the data, see Figure 1 that shows that the pilot smoother (plain line) is nearly constant. The iterative bias corrected estimators are plotted in figure (1) for values of k , the number of iterations, in $\{1, 10, 50, 100, 500, 10^3, 10^5, 10^6\}$

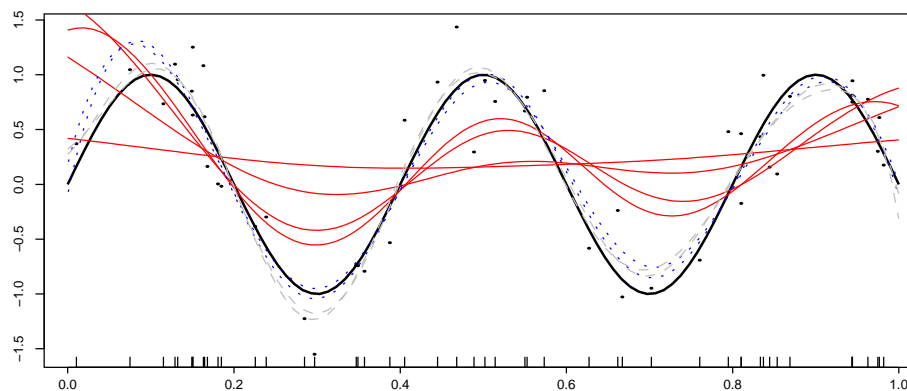


FIG 1. True function m_1 (fat plain line) and different estimators varying with the number of iterations k .

Figure 1 shows how each bias correction iteration changes the smoother, starting from a nearly constant smoother and slowly deforming (going down into the valleys and up into the peaks) with increasing number of iterations $k = 10$, $k = 50$ and $k = 100$. After 500 iterations, the iterative smoother is very close to the true function. However when the number of iterations is very large (here $k = 10^5$ and 10^6) the iterative smoother deteriorates.

LEMMA 2.2. The squared bias and variance of the k^{th} iterated bias corrected linear smoother \hat{m}_k (2.4) are

$$\begin{aligned} B^2(\hat{m}_k) &= m^t \left((I - S)^k \right)^t (I - S)^k m \\ V(\hat{m}_k) &= \sigma^2 (I - (I - S)^k) \left((I - (I - S)^k) \right)^t. \end{aligned}$$

Remark: Symmetric smoothing matrices S can be decomposed as $S = P_S \Lambda_S P_S^t$, with orthonormal matrix $P_S = [u_1, u_2, \dots, u_n]$ and diagonal matrix Λ_S .

$$(2.6) \quad \hat{m}_k = P_S \text{diag}(1 - (1 - \Lambda_S)^k) P_S^t Y = \sum_j (1 - (1 - \lambda_j)^k) u_j u_j^t Y.$$

Applying Lemma 2.2, we get

$$\begin{aligned} B^2(\hat{m}_k) &= m^t P_S (I - \Lambda_S)^{2k} P_S^t m \\ V(\hat{m}_k) &= \sigma^2 P_S (I - (I - \Lambda_S)^k)^2 P_S^t. \end{aligned}$$

Hence if the magnitude of the eigenvalues of $I - S$ are bounded by one, each iteration of the bias correction will decrease the bias and increase the variance. This monotonicity (decreasing bias, increasing variance) with increasing number of iterations k allows us consider data driven selection for number of bias correction steps to achieves the best compromise between bias and variance of the smoother.

The preceding remark suggests that the behavior of the iterative bias corrected smoother \hat{m} is tied to the spectrum of $I - S$, and not of S . The next theorem collects the various convergence results for iterated bias corrected linear smoothers.

THEOREM 2.3. *Suppose that the singular values $\lambda_j = \lambda_j(I - S)$ of $I - S$ satisfy*

$$(2.7) \quad -1 < \lambda_j < 1 \quad \text{for } j = 1, \dots, n.$$

Then we have that

$$\begin{aligned} \|\hat{b}_k\| &< \|\hat{b}_{k-1}\| \quad \text{and} \quad \lim_{k \rightarrow \infty} \hat{b}_k = 0, \\ \|R_k\| &< \|R_{k-1}\| \quad \text{and} \quad \lim_{k \rightarrow \infty} R_k = 0, \\ \lim_{k \rightarrow \infty} \hat{m}_k &= Y \quad \text{and} \quad \lim_{k \rightarrow \infty} \mathbb{E}[\|\hat{m}_k - m\|^2] = n\sigma^2. \end{aligned}$$

Conversely, if $I - S$ has a singular value $|\lambda_j| > 1$, then

$$\lim_{k \rightarrow \infty} \|\hat{b}_k\| = \lim_{k \rightarrow \infty} \|R_k\| = \lim_{k \rightarrow \infty} \|\hat{m}_k\| = \infty.$$

Remark 1: This theorem shows that iterating the booting algorithm to reach the limit of the sequence of boosted smoothers, Y_∞ , is not the desirable.

However, since each iteration decreases the bias and increases the variance, a suitably stopped Boosting estimator is likely to improve upon the initial smoother.

Remark 2: When $|\lambda_j(I - S)| > 1$, the iterative bias correction fails. The reason is that \hat{b}_k overestimates the true bias b_k , and hence Boosting repeatedly overcorrects the bias of the smoothers, which results in a divergent sequence of smoothers. Divergence of the sequence of boosted smoothers can be detected numerically, making it possible to avoid this bad behavior by combining the iterative bias correction procedure with a suitable stopping rule.

Remark 3: The assumption that for all j , the singular values $-1 < \lambda_j(I - S) < 1$ implies that $I - S$ is a contraction, so that $\|(I - S)Y\| < \|Y\|$. This condition does not imply that the smoother S itself is a shrinkage smoother as defined by (Buja et al. [5]). Conversely, not all shrinkage estimators satisfy the condition 2.7 of the theorem. In Section 3, we will give examples of common shrinkage smoothers for which $|\lambda_j(I - S)| > 1$, and show numerically that for these shrinkage smoothers, the iterative bias correction scheme will fail.

2.2. L_2 Boosting for regression. Boosting is one of the most successful and practical methods that arose 15 years ago from the machine learning community (Freund [15], Schapire [31]). In light of Friedman [16], the Boosting algorithms has been interpreted as functional gradient descent technique. Let us summarize the L_2 Boost algorithm described in Buhlmann and Yu [4].

Step 0: Set $k = 1$. Given the data $\{(X_i, Y_i), i = 1, \dots, n\}$, calculate an pilot regression smoother

$$\hat{F}_1(x) = h(x; \hat{\theta}_{X,Y}),$$

by least squares fitting of the parameter, that is,

$$\hat{\theta}_{X,Y} = \operatorname{argmin}_{\theta} \sum_{i=1}^n (Y_i - h(X_i, \theta))^2.$$

Step 1: With a current smoother \hat{F}_k , compute the residuals $U_i = Y_i - \hat{F}_k(X_i)$ and fit the real-valued learner to the current residuals by least square. The fit is denoted by $\hat{f}_{k+1}(\cdot)$. Update

$$(2.8) \quad \hat{F}_{k+1}(\cdot) = \hat{F}_k(\cdot) + \hat{f}_{k+1}(\cdot).$$

Step 2: Increase iteration index k by one and repeat step 1.

LEMMA 2.4 (Buhlmann and Yu, 2003). *The smoothing matrix associated with the k^{th} Boosting iterate of linear smoother with smoothing matrix S is*

$$\hat{F}_k = (I - (I - S)^k)Y = B_k Y.$$

Viewing Boosting as a greedy gradient descent method, the update formula (2.8) is often modified to include *convergence factor* μ_k , as in Friedman [16], to become

$$\hat{F}_{k+1}(\cdot) = \hat{F}_k(\cdot) + \hat{\mu}_{k+1} \hat{f}_{k+1}(\cdot),$$

where $\hat{\mu}_{k+1}$ is the best step toward the best direction $\hat{f}_{k+1}(\cdot)$.

This general formulation allows a great deal of flexibility, both in selecting the type of smoother used in each iteration of the Boosting algorithm, and in the selection of the convergence factor. For example, we may start with a running mean pilot smoother, and use a smoothing spline to estimate the bias in the first Boosting iteration and a nearest neighbor smoother to estimate the bias in the second iteration. However in practice, one typically uses the same smoother for all iterations and fix the convergence factor $\mu_k \equiv \mu \in (0, 1)$. That is, the sequence of smoothers resulting from the Boosting algorithm is given by

$$(2.9) \quad \hat{F}_k = (I - (I - \mu S)^k)Y = B_k Y.$$

We shall discuss in detail in Section 4 the impact of this convergence factor and other modifications to the Boosting algorithm to ensure good behavior of the sequence of boosted smoothers.

3. Boosting classical smoothers. This section is devoted to understanding the behavior of the iterative Boosting schema using classical smoothers, which in light of Theorem 2.3, depends on the magnitude of the singular values of the matrix $I - S$.

We start our discussion by noting that Boosting a **projection type smoothers** is of no interest because residuals $(I - S)Y$ are orthogonal to smoother SY . It follows that the smoothed residuals $S(I - S)Y = 0$, and as a result, $\hat{m}_k = \hat{m}_1$ for all k . Hence Boosting a bin smoother or a regression spline smoother leaves the initial smoother unchanged.

Consider the **K -nearest neighbor smoother**. Its associated smoothing matrix is $S_{ij} = 1/K$ if X_j belongs to the K -nearest neighbor of X_i and

$S_{ij} = 0$ otherwise. Note that this smoothing matrix is not symmetric. While this smoother enjoys many desirable properties, it is not well suited for Boosting because the matrix $I - S$ has singular values larger than one.

THEOREM 3.1. *In the fixed design or in the uniform design, as soon as the number of K is bigger than one and smaller than n , at least one singular value of $I - S$ is bigger than 1.*

The proof of the theorem is found in the appendix. A consequence of Proposition 3.1 and Theorem 2.3, is that the Boosting algorithm applied to a K -nearest neighbor smoother produces a sequence of divergent smoothers, and hence should not be used in practice.

Example continued with K -nearest neighbor smoother. We confirm this behavior numerically. Using the same data as before, we apply the Boosting algorithm starting with an pilot K -nearest neighbor smoother with $K = 10$. The pilot estimator is plotted in a plain line, and the various boosted smoothers with k , the number of iterations, valued in $\{2, \dots, 5\}$ in dotted lines.

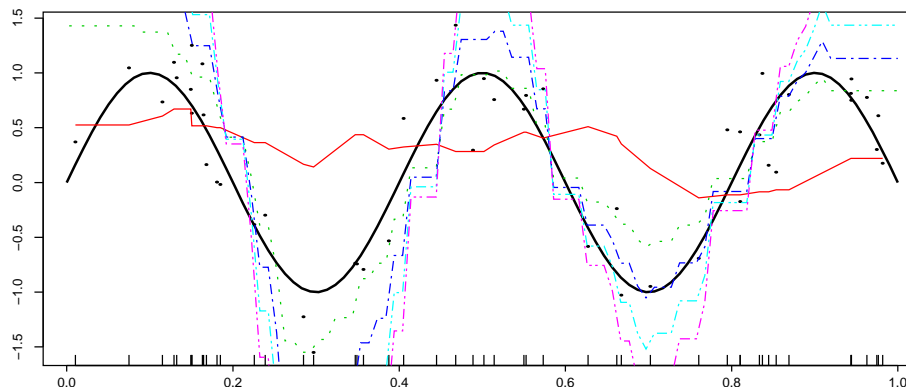


FIG 2. True function m_1 (fat plain line) and different estimators varying with the number of iterations k .

For $k = 1$, the pilot smoother is nearly constant (since we take $K = 10$ neighbors) and very quickly the iterative bias corrected estimator explodes. Qualitatively, the smoothers are getting higher at the peaks and lower in the valleys, which is consistent with an overcorrection of the bias. Contrast this behavior with the one shown in Figure 1.

Kernel type smoother. For Nadaraya kernel type estimator, the smoothing matrix S has entries $S_{ij} = K_h(X_i - X_j) / \sum_k K_h(X_i - X_k)$, where $K(\cdot)$ is a symmetric function (e.g., uniform, Epanechnikov, Gaussian), h denotes the bandwidth and $K_h(\cdot)$ is the scaled kernel $K_h(t) = h^{-1}K(t/h)$. The matrix S is not symmetric but can be written as $S = D\mathbb{K}$ where \mathbb{K} is symmetric with general element $[K_h(X_i - X_j)]$ and D is diagonal with element $1 / \sum_j K_h(X_i - X_j)$. Algebraic manipulations allows us to rewrite the iterated estimator as

$$\begin{aligned}\hat{m}_k &= [I - (I - S)^k]Y \\ &= [I - (D^{1/2}D^{-1/2} - D^{1/2}D^{1/2}\mathbb{K}D^{1/2}D^{-1/2})^k]Y \\ &= [I - D^{1/2}(I - D^{1/2}\mathbb{K}D^{1/2})^kD^{-1/2}]Y \\ &= D^{1/2}[I - (I - A)^k]D^{-1/2}Y.\end{aligned}$$

Since the matrix $A = D^{1/2}\mathbb{K}D^{1/2}$ is symmetric, we apply the classical decomposition $A = P_A\Lambda_AP_A^t$, with P_A orthonormal and Λ_A diagonal, to get a closed form expression for the boosted smoother

$$\hat{m}_k = D^{1/2}P_A[I - (I - \Lambda_A)^k]P_A^tD^{-1/2}Y.$$

The eigen decomposition of $A = D^{1/2}\mathbb{K}D^{1/2}$ can be used to describe the behavior of the sequence of iterative estimators. In particular, any eigenvalue of $A = D^{1/2}\mathbb{K}D^{1/2}$ that is negative or greater than 2 will lead to unstable procedure. If the kernel $K(\cdot)$ is a symmetric probability density function positive definite, then the spectrum of the Nadaraya-Watson kernel smoother lies between zero and one.

THEOREM 3.2. *If the inverse Fourier-Stieltjes transform of a kernel $K(\cdot)$ is a real positive finite measure, then the spectrum of the Nadaraya-Watson kernel smoother lies between zero and one.*

Conversely, suppose that X_1, \dots, X_n are an independent n -sample from a density f (with respect to Lebesgue measure) that is bounded away from zero on a compact set strictly included in the support of f . If the inverse Fourier-Stieltjes transform of a kernel $K(\cdot)$ is not a positive finite measure, then with probability approaching one as the sample size n grows to infinity, the maximum of the spectrum of $I - S$ is larger than one.

Remark 1: Since the $\text{spec}(A)$ is the same as the $\text{spec}(S)$ and S is row stochastic, we conclude that $\text{spec}(A) \leq 1$. So we are only concern by the presence of negative eigenvalues in the spectrum of A .

Remark 2: In Di Marzio and Taylor [10] proved the first part of the theorem. Our proof of the converse shows that for large enough sample sizes,

most configurations from a random design lead to smoothing matrix S with negative singular values.

Remark 3: The assumption that the inverse Fourier-Stieltjes transform of a kernel $K(\cdot)$ is a real positive finite measure is equivalent to the kernel $K(\cdot)$ being positive a definite function, that is, for any finite set of points x_1, \dots, x_m , the matrix

$$\begin{pmatrix} K(0) & K(x_1 - x_2) & K(x_1 - x_3) & \dots & K(x_1 - x_m) \\ K(x_2 - x_1) & K(0) & K(x_2 - x_3) & \dots & K(x_2 - x_m) \\ \vdots & & & & \vdots \\ K(x_m - x_1) & K(x_m - x_2) & K(x_m - x_3) & \dots & K(0) \end{pmatrix}$$

is positive definite. We refer to Schwartz [32] for a detailed study of positive definite functions.

The Gaussian and triangular kernels are positive definite kernels (they are the Fourier transform of a finite positive measure (Feller [14])) and in light of Theorem 3.2 the Boosting of Nadaraya-Watson kernel smoothers with these kernels produces a sequence of well behavior smoother. However, the uniform and the Epanechnikov kernels are not positive definite. Theorem 3.2 states that for large samples, the spectrum of $I - S$ is larger than one and as a result the sequence of boosted smoother diverges. Proposition 3.3 below strengthen this result by stating that the largest singular value of $I - S$ is always larger than one.

PROPOSITION 3.3. *Let S be the smoothing matrix of a Nadaraya-Watson regression smoother based on either the uniform or the Epanechnikov kernel. Then the largest singular value of $I - S$ is larger than one.*

Example continued with Epanechnikov kernel smoother. In the next figure, the pilot smoother is a kernel one with an Epanechnikov kernel and with bandwidth is equal to 0.15. The pilot smoother is the plain line, and the subsequent iterations with k , the number of iterations, valued in $\{1, 2, 5, 10, 20, 50, 100\}$, are the dotted lines.

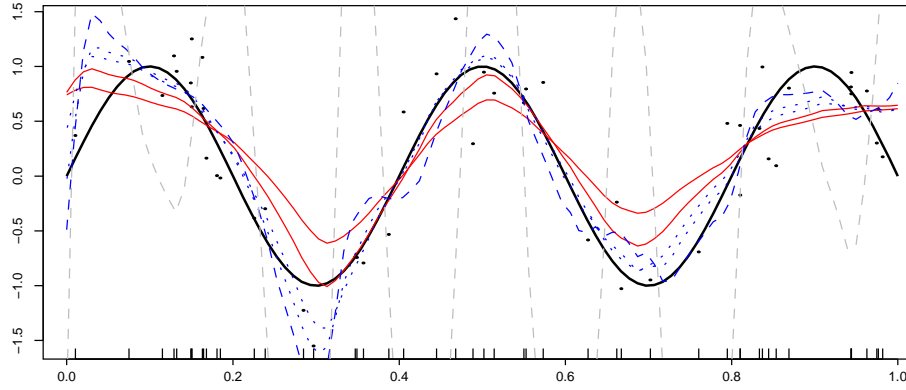


FIG 3. True function m_1 (fat plain line) and different estimators varying with the number of iterations k .

For $k = 1$, the pilot smoother oversmooths the true regression since the bandwidth takes almost one third of the data and very quickly the iterative estimator explodes. Contrast this behavior with the one shown by the Gaussian kernel smoother in Figure 1.

Finally, let us now consider the **smoothing splines smoother**. The smoothing matrix S is symmetric, and therefore admits an eigen decomposition. Denote by $\{u_1, u_2, \dots, u_n\}$ an orthonormal basis of eigenvectors of S associated to the eigenvalues $1 \geq \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$ (Utreras [36]). Denote by $P_S = [u_1, u_2, \dots, u_n]$ the orthonormal matrix of column eigenvectors and write $S = P_S \text{diag}(\lambda_S) P_S^t$, that is $S = \sum_j \lambda_j u_j u_j^t$. The iterated bias reduction estimator is given by (2.6). Since all the eigenvalues are between 0 and 1, then if k is large, the iterative procedure kills the eigenvalues less than 1 and put the others to 1.

Example continued with smoothing splines smoother In the next figure, the pilot smoother is a smoothing spline, with λ equals to 0.2. The different estimators are plotted in figure (4), with the pilot estimator in plain line and the boosted smoothers with number of iterations k being $\{10, 50, 100, 500, 10^3, 10^5, 10^6\}$ in dotted lines.

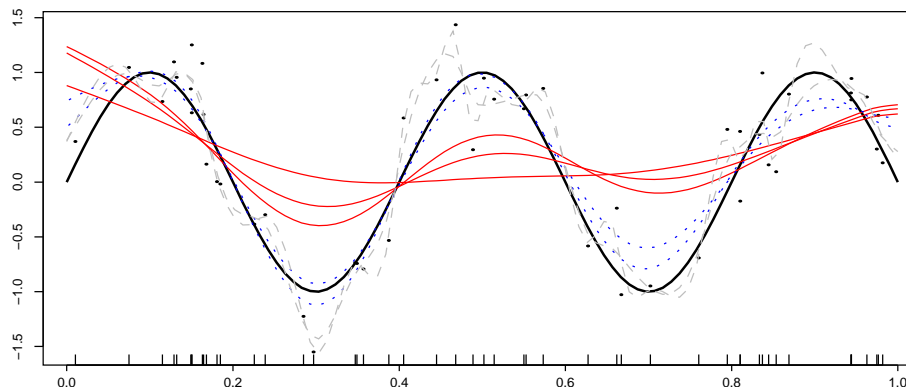


FIG 4. True function m_1 (fat plain line) and different estimators varying with the number of iterations k .

The pilot estimator is more variable than the pilot estimator of figure 1 and by the way the convergence and the deterioration arise faster.

4. Smoother engineering. Practical implementations of the Boosting algorithm include a user selected convergence factor $\mu \in (0, 1)$ that appears in the definition of the boosted smoother

$$(4.1) \quad \hat{m}_k = (I - (I - \mu S)^k)Y = B_k Y.$$

In this section, we show that when $\mu < 1$, one effectively operates a partial bias correction. This partial bias correction does not however resolve the problems associated with Boosting a nearest neighbor or Nadaraya Watson kernel smoothers with compact kernel we exhibited in the previous section. To resolve these problems, we propose to suitably modify the boosted smoother. We call such targeted changes *smoother engineering*.

The following iterative partial bias reduction scheme is equivalent to the Boosting algorithm defined by Equation (4.1): Given a smoother $\hat{m}_k = B_k Y$ at the k^{th} iteration of the Boosting algorithm, calculate the residuals R_k and estimated bias \hat{b}_k

$$\begin{aligned} R_k &= Y - \hat{m}_k = (I - B_k)Y \\ \hat{b}_k &= S R_k = S(I - B_k)Y. \end{aligned}$$

Next, given $0 < \mu < 1$, consider the partially bias corrected smoother

$$(4.2) \quad \hat{m}_{k+1} = \hat{m}_k + \mu \hat{b}_k.$$

Algebraic manipulations of the smoothing matrix of the right-hand side of (4.2) yields

$$B_k + \mu S(I - B_k) = I - (I - \mu S)^{k+1},$$

from which we conclude that \hat{m}_{k+1} satisfies (4.1) and therefore is the $(k+1)^{th}$ iteration of the Boosting algorithm. It is interesting to rewrite (4.2) as

$$\hat{m}_{k+1} = (1 - \mu)\hat{m}_k + \mu [\hat{m}_k + \hat{b}_k],$$

which shows that boosted smoother \hat{m}_{k+1} is a convex combination between the smoother \hat{m}_k at iteration k , and the fully bias corrected smoother $\hat{m}_k + \hat{b}_k$. As a result, we understand how the introduction of a convergence factor produces a "weaker learner" than the one obtained for $\mu = 1$.

In analogy to Theorem 2.3, the behavior of the sequence of the smoother depends on the spectrum of $I - \mu S$. Specifically, if $\max_j |\lambda_j(I - \mu S)| \leq 1$, then $\lim_{k \rightarrow \infty} \hat{m}_k = Y$, and conversely, if $\max_j |\lambda_j(I - \mu S)| > 1$, $\lim_{k \rightarrow \infty} \|\hat{m}_k\| = \infty$. Inspection of the proofs of propositions 3.1 and 3.2 reveal that the spectrum of $(I - \mu S)$ for both the nearest neighbor smoother and the Nadaraya Watson kernel smoother has singular values of magnitude larger than one. Hence the introduction of the convergence factor does not help resolve the difficulties arising when Boosting these smoothers.

To resolve the potential convergence issues, one needs to suitably modify the underlying smoother to ensure that the magnitude of the singular values of $I - \mu S$ are bounded by one. A practical solution is to replace the smoothing matrix S by $S^* = SS^t$. If S is a contraction, it follows that the eigenvalues of $I - S^*$ are nonnegative and bounded by one. Hence the Boosting algorithm with this smoother will produce a well behaved sequence of smoothers with $\lim_{k \rightarrow \infty} \hat{m}_k = Y$.

While substituting the smoother S^* for S can produce better boosted smoothers in cases where Boosting failed, our numerical experimentations has shown that the performance of Boosting S^* is not as good as Boosting S when the pilot estimator enjoyed good properties, as is the case for smoothing splines and the Nadaraya Watson kernel smoother with Gaussian kernel.

5. Stopping rules. Theorem 2.3 in Section 2 states that the limit of the sequence of boosted smoothers is either the raw data Y or has norm $\|Y_\infty\| = \infty$. It follows that iterating the Boosting algorithm until convergence is not desirable. However, since each iteration of the Boosting algorithm reduces the bias and increases the variance, often a few iteration of the Boosting

algorithm will produce a better smoother than the pilot smoother. This brings up the important question of how to decide when to stop the iterative bias correction process.

Viewing the latter question as a model selection problem suggests stopping rules based on Mallows' C_p (Mallows [28]), Akaike Information Criterion, AIC, (Akaike [1]), Bayesian Information Criterion, BIC, (Schwarz [33]), and Generalized cross validation (Craven and Wahba [7]). Each of these selectors estimate the optimum number of iterations k of the Boosting algorithm by minimizing estimates of the expected squared prediction error of the smoothers over some pre-specified set $\mathcal{K} = \{1, 2, \dots, M\}$.

Three of the six criteria we study numerically in Section 6 use plug-in estimates for the squared bias and variance of the expected prediction mean square error. Specifically, consider

$$(5.1) \quad \hat{k}_{AIC} = \operatorname{argmin}_{k \in \mathcal{K}} \left\{ \hat{\sigma}^2 + 2 \frac{\operatorname{trace}(S_k)}{n} \right\},$$

$$(5.2) \quad \hat{k}_{GCV} = \operatorname{argmin}_{k \in \mathcal{K}} \left\{ \log \hat{\sigma}^2 - 2 \log \left(1 - \frac{\operatorname{trace}(S_k)}{n} \right) \right\},$$

$$(5.3) \quad \hat{k}_{AIC_C} = \operatorname{argmin}_{k \in \mathcal{K}} \left\{ \log \hat{\sigma}^2 + 1 + \frac{2(\operatorname{trace}(S_k) + 1)}{n - \operatorname{trace}(S_k) - 2} \right\}.$$

In nonparametric smoothing, the AIC criteria (5.1) has a noticeable tendency to select more iterations than needed, leading to a final smoother $\hat{m}_{\hat{k}_{AIC}}$ that typically undersmooths the data. As a remedy, Hurvich et al. [24] introduced a corrected version of the AIC (5.3) under the simplifying assumption that the nonparametric smoother \hat{m} is unbiased, which is rarely hold in practice and which is particularly not true in our context.

The other three criteria considered in our simulation study in Section 6 are Cross-Validation, L-fold cross-validation and data splitting, all of which estimate empirically the expected prediction mean square error by splitting the data into learning and testing sets. Implementation of these criterion require one to evaluate the smoother at locations outside the of the design. To this end, write the k^{th} iterated smoother as a k times bias corrected smoother

$$\begin{aligned} \hat{m}_k &= \hat{m}_0 + \hat{b}_1 + \dots + \hat{b}_k \\ &= S[I + (I - S) + (I - S)^2 + \dots + (I - S)^{k-1}]Y, \end{aligned}$$

which we rewrite as

$$\hat{m}_k = S\hat{\beta}_k,$$

where

$$\begin{aligned}\hat{\beta}_k &= [I + (I - S) + (I - S)^2 + \cdots + (I - S)^{k-1}]Y \\ &= Y + R_1 + R_2 + \cdots + R_k\end{aligned}$$

is a vector of size n . Given the vector $S(x)$ of size n whose entries are the weights for predicting $m(x)$, we calculate

$$(5.4) \quad \hat{m}_k(x) = S(x)^t \hat{\beta}_k.$$

This formulation is computationally advantageous because the vector of weights $S(x)$ only needs to be computed once, while each Boosting iteration updates the parameter vector $\hat{\beta}_k$ by adding the residuals $R_k = Y - \hat{m}_k$ of the fit to the previous value of the parameter, i.e., $\hat{\beta}_k = \hat{\beta}_{k-1} + R_k$. The vector $S(x)$ is readily computed for many of the smoothers used in practice. For kernel smoothers, the i^{th} entry in the vector $S(x)$ is

$$S_i(x) = \frac{K_h(x - X_i)}{\sum_j K_h(x - X_j)}.$$

For smoothing spline, let $N(x)$ denote the vector of basis function evaluated at x . One can show that $\hat{m}_k(x) = N(x)M\hat{\beta}_k$, where M is the $n \times n$ matrix given by

$$M = (N^t N + \lambda \Omega)^{-1} N^t.$$

Finally, for the K -nn smoother, the entries of the vector $S(x)$ are

$$S_i(x) = \begin{cases} 1/K & \text{if } X_i \text{ is a } K\text{-nn of } x \\ 0 & \text{otherwise} \end{cases}.$$

We note that if the spectrum of $I - S$ is bounded in absolute value by one, then the parameter $\hat{\beta}_k \rightarrow \beta_\infty$, and hence we have pointwise convergence of $\hat{m}_k(x)$ to some $m_\infty(x)$, whose properties depend on $S(x)$.

To define the data splitting and cross validation stopping rules, one divides the sample into two disjoint subset: a learning set \mathcal{L} which is used to estimate the smoother $\hat{m}^\mathcal{L}$, and a testing set \mathcal{T} on which predictions from the smoother are compared to the observations. The data splitting selector for the number of iterations is

$$(5.5) \quad \hat{k}_{DS} = \operatorname{argmin}_{k \in \mathcal{K}} \sum_{i \in \mathcal{T}} \left(Y_i - \hat{m}_k^\mathcal{L}(X_i) \right)^2.$$

One-fold cross validation, or simply cross validation, and more generally L -fold cross validation average the prediction error over all partitions of the data into learning and testing sets, with fixed size of the testing set $|\mathcal{T}| = L$. This leads to

$$(5.6) \quad \hat{k}_{CV} = \operatorname{argmin}_{k \in \mathcal{K}} \sum_{|\mathcal{T}|=L} \sum_{i \in \mathcal{T}}^n \left(Y_i - \hat{m}_k^{\mathcal{L}}(X_i) \right)^2.$$

We rely on the expansive literature on model selection to provide insight into the statistical properties of stopped boosted smoother. For example, Theorem 3.2 of Li [27] describes the asymptotic behavior of the generalized cross-validation (GCV) stopping rule applied to spline smoothers.

THEOREM 5.1 (Li, 1987). *Assume that Li's assumptions are verified for the smoother S . Then*

$$\frac{\|m - S_{\hat{k}_{GCV}} Y\|^2}{\inf_{k \in \mathcal{K}} \|m - S_k Y\|^2} \rightarrow 1 \quad \text{in probability.}$$

Results on the finite sample performance for data splitting for arbitrary smoothers is presented in Theorem 1 of Hengartner et al. [21] who proved the following oracle inequality.

THEOREM 5.2. *For each k in \mathcal{K} , $\lambda > 0$ and $\alpha > 0$, we have*

$$\begin{aligned} P \left\{ \frac{1}{m} \sum_{i=n+1}^{n+m} (\hat{m}_{K_{DS}} - m)^2(X_i) - (1 - \alpha) \sum_{i=n+1}^{n+m} (\hat{m}_k - m)^2(X_i) \geq \lambda \right\} \\ \leq |K| \sqrt{\left(\frac{32(1 + \alpha)\sigma^2}{\pi \alpha m \lambda} \right) \left[\exp \left(\frac{\alpha m \lambda}{8(1 + \alpha)\sigma^2} \right) - 1 \right]^{-1}}. \end{aligned}$$

Example continued with smoothing splines

Figure 5 shows the three pilot smoothers (smoothing splines with different smoothing parameters) considered in the simulation study in Section 6.

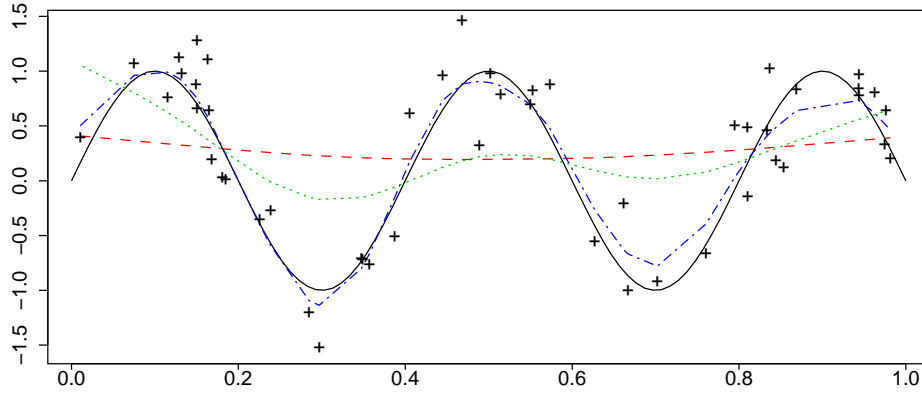


FIG 5. True function m_1 (plain line) and different pilot smoothing spline smoother, $S(\lambda_1)$ (dotted line), $S(\lambda_2)$ (dashed line), $S(\lambda_3)$ (dash-dotted line) for the 50 data points of one replication (Gaussian error).

Starting with the smoothest pilot smoother $S(\lambda_1)$, the Generalized Cross Validation criteria stops after 1389 iterations. Starting with smoother $S(\lambda_2)$, GCV stopped after 23 iterations, while starting with the noisiest pilot $S(\lambda_3)$, GCV stopped after one iteration. It is remarkable how similar the final smoother are.

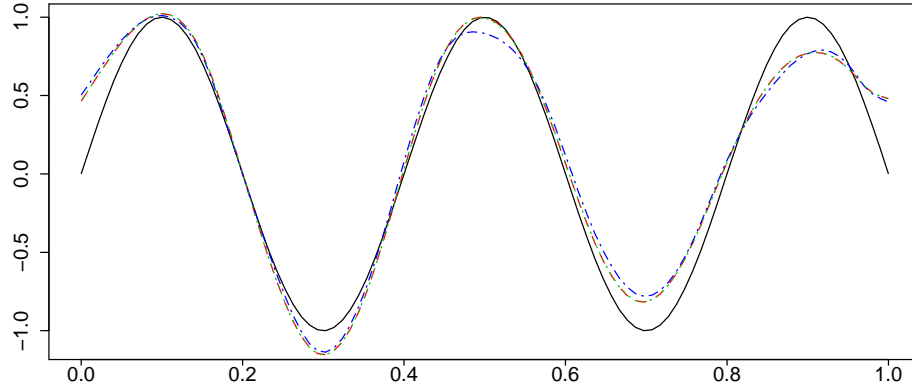


FIG 6. True function m_1 (plain line) and different pilot smoothing spline smoother, $S(\lambda_1)$ (dashed line), $S(\lambda_2)$ (dotted line), $S(\lambda_3)$ (dash-dotted line) for the same 50 data points as in figure 5 of one replication (Gaussian error).

The final selected estimators are very close to one another, despite the different pilot smoothers and the different numbers iterations that were selected by the GVC criteria. Despite the flatness of the pilot smoother $S(\lambda_1)$, it succeeds after 1389 iteration to capture the signal. Note that larger smoothing

parameter λ are associated to weaker learners that require a larger number of bias correction iterations before they become desirable smoothers according to the generalized cross validation criteria. A close examination of figure 6 shows that using the less biased estimator $S(\lambda_3)$ leads to the worse final estimator. This can be explained as follows: if the pilot smoother is not enough biased, after the first step almost no signal is left in the residuals and the iterative bias reduction is stopped.

We remark again that one does not need to keep the same smoother throughout the iterative bias correcting scheme. We conjecture that there are advantages to using weaker smoothers later in the iterative scheme, and shall investigate this in a forthcoming paper.

6. Simulations. This section presents selected results from our simulation study that investigates the statistical properties of the various data driven stopping rules. The simulations examine, within the framework set by Hurvich et al. [24], the impact on performance of various stopping rules, smoother type, smoothness of the pilot smoother, sample size, true regression function, and the relative variance of the errors as measured by the signal to noise ratio.

We examine the influence of various factors on the performance of the selectors, with 100 simulation replications and a random uniform grid in $[0, 1]$. The error standard deviation is $\sigma = 0.2R_g$, where R_g is the range of $g(x)$ over $x \in [0, 1]$. For each setting of factors, we have

- (A) sample size: $n = 50, 100$ and 500
- (B) the following 3 regression functions, most of which were used in earlier studies
 1. $m(x) = \sin(5\pi x)$,
 2. $m(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$,
 3. $m(x) = \exp(x - \frac{1}{3})\{x < \frac{1}{3}\} + \exp[-2(x - \frac{1}{3})]\{x \geq \frac{1}{3}\}$.
- (C) error distribution: Gaussian and Student(5)
- (D) pilot smoothers: smoothing splines, Gaussian kernel, K -nearest neighbor type smoother
- (E) three starting smoothers: S_1, S_2 and S_3 by decreasing order of smoothing.

For each setting, we compute the ideal numbers of iterations by computing at data points $\{X_i\}_{i=1}^n$

$$k_{\text{opt}} = \underset{k \in \mathcal{K}}{\operatorname{argmin}} \sum_{i=1}^n \|m(X_i) - \hat{m}_k(X_i)\|^2.$$

Since the results are numerous we report here a summary, focusing on the main objectives of the paper.

First of all, does the stopping procedures proposed in section 5 work ? Figure 7 represents the kernel density estimates of the log ratios of the number of iterations to the ideal number of iterations for the smoothing spline type smoother.

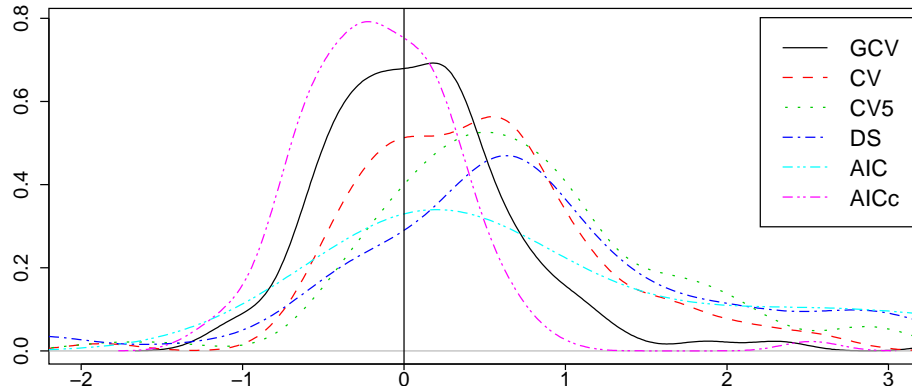


FIG 7. *Estimated density of $\log(\hat{k}/k_{\text{opt}})$, \hat{k} evaluated by different stopping criterion : GCV, CV (leave one out), CV 5 fold (CV5), data splitting (DS), AIC and corrected AIC (AICc). Density is estimated on 100 replication for function m_1 , with Gaussian error, spline smoother S_2 and $n = 50$ data points.*

Obviously, negative values indicate undersmoothing (\hat{k} smaller than k_{opt} , that is not enough bias reduction) while positive values indicate oversmoothing. The results remain essentially unchanged over the range of starting values, regression function and smoothers types we considered in our simulation study.

For small data sets ($n = 50$), the stopping rule based on data splitting produced values for \hat{k} that were very variable. A similar observation about the variability of bandwidth selection from data splitting was made in [see 21]. We also found that the five fold cross validation stopping rule produced highly variable values for \hat{k} .

The AIC stopping rule selects values \hat{k} that are often too big (oversmoothing) and sometimes selects the largest possible value of $\hat{k} \in K$. In that cases, the curve k versus AIC (not shown) indicates two minimum, a local one which is around the true value and the global one at the boundary. This can be attributed to the fact that the penalty term used by AIC is too small. The AICc criteria uses a larger penalty term, which leads to smaller values

for \hat{k} . In fact, the selected values are typically smaller than the optimal one. The penalty associated with GCV lies in between the AICc penalty and AIC penalty, and produces in practice values of \hat{k} that are closer to the optimum than either AIC or AICc. Finally, the leave one out cross-validation selection rule produces \hat{k} that are typically larger than the optimal one.

Investigation of the MSE as a function of the number of iterations k reveal that, in the examples we considered, that function decreases rapidly towards its minimum and then remains relatively flat over a range of values to right of the minimum. It follows that the loss of stopping after k_{opt} is less than stopping before k_{opt} . We verify this empirically as follows: for each estimate, we calculate the approximation to the integrated mean squared error between the estimator $\hat{m}_{\hat{k}}$ and the true regression function m

$$\text{MSE}(\hat{m}_{\hat{k}}) = 1/100 \sum_{x \in \mathcal{G}} |m(x) - \hat{m}_{\hat{k}}(x)|^2,$$

where \mathcal{G} is a fix grid of 100 regularly spaced points in the unit interval $[0, 1]$. We partition the calculated integrated mean squared error depending on whether \hat{k} is bigger than k_{opt} or smaller than k_{opt} . Figure 8 presents the boxplot of the integrated mean squared error when \hat{k} over-estimates k_{opt} and when it under-estimates k_{opt} and clearly shows that over-estimating k_{opt} leads to smaller integrated mean squared error than under-estimating k_{opt} .

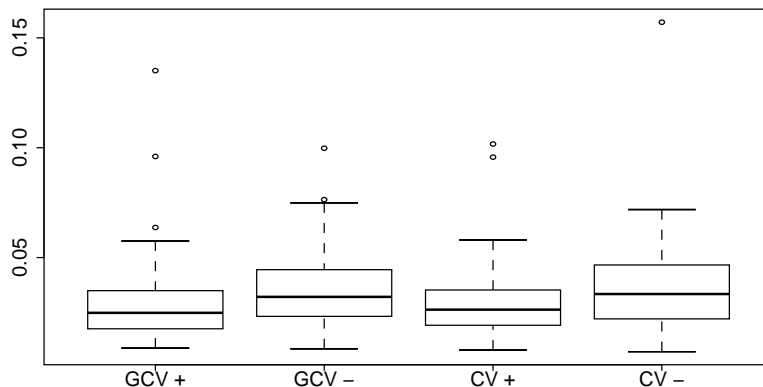


FIG 8. Boxplot of $\text{MSE}_{\hat{m}_{\hat{k}}}$ when $\hat{k}_{\text{GCV}} > k_{\text{opt}}$ (denoted as GCV+), of mean squared error of $\hat{f}_{\hat{k}_{\text{GCV}}}$ when $\hat{k}_{\text{GCV}} \leq k_{\text{opt}}$ (denoted as GCV-), and the same boxplots with leave one out stopping criterion. Mean squared error are estimated on a grid of 100 points regularly spaced between 0 and 1, 100 replication for function m_1 , with Gaussian error, spline smoother S_2 and $n = 50$ data points.

For bigger data sets, say $n = 100$ or bigger, most of the stopping criterion act the same except for the modified AIC which tends to select a smaller number of iterations k than the ideal one. One fold cross-validation is rather computational intensive as the usual relation between cross validated estimator at X_i and full data estimator is no longer valid [e.g. 19, p. 47].

These conclusions remain true for kernel smoother and nearest neighbor smoothers. However if the pilot smoother is not smooth enough (not biased enough), then the number of iteration is too small to allow us to discriminate between the different stopping rules. These initial smoothers we name as wiggly learner are almost unbiased and therefore, there is little value to apply an iterative bias correction scheme. In conclusion, for small data sets, our simulations show that both GCV and leave one cross-validation work well, and for bigger data sets, we recommend using GCV.

Tables (1) and (2) here below report the finite sample performance of stopped boosted smoother by the GCV criterion. Each entry in the table reports the median number of iterations and the median mean square error over 100 simulations. As expected, larger smoothing parameter of the initial smoother require more iterations of the boosting algorithm to reach its optimum. Interestingly, the selected smoother starting with a very smooth smoother, has slightly smaller mean squared error. To quantify the benefits of the iterative bias correction scheme, the last column of the tables gives the mean squared error of the original smoother with smoothing parameters selected using GCV. In all cases, the iterative bias correction has smaller mean squared error than the "one-step" smoother, with improvements ranging from 15% to 30%.

Table (1) presents the results for smoothing splines.

Function m_1							
error	\hat{k}_1	$S_{\hat{k}_{\text{GCV}}}(\lambda_1)$	\hat{k}_2	$S_{\hat{k}_{\text{GCV}}}(\lambda_2)$	\hat{k}_3	$S_{\hat{k}_{\text{GCV}}}(\lambda_3)$	$S(\hat{\lambda}_{\text{GCV}})$
Gaussian	4077	0.0273	65	0.0282	2	0.0293	0.0379
student	4115	0.0273	70	0.0286	2	0.0296	0.0352
Function m_2							
Gaussian	1219	0.0798	21	0.0845	1	0.0837	0.0829
student	1307	0.0887	22	0.0944	1	0.0932	0.0937
Function m_3							
Gaussian	135	0.0014	3	0.0014	1	0.0016	0.0016
student	147	0.0016	3	0.0016	1	0.0018	0.0019

TABLE 1

Median over 100 simulations of the number of iterations and the MSE for smoothing splines smoother, $n = 50$ data points. $S(\hat{\lambda}_{\text{GCV}})$ denotes the traditional smoothing splines estimate with λ chosen with GCV.

Table (2) presents the results for kernel smoothers with a Gaussian kernel.

Function m_1							
error	\hat{k}_1	$S_{\hat{k}_{\text{GCV}}}(h_1)$	\hat{k}_2	$S_{\hat{k}_{\text{F-GCV}}}(h_2)$	\hat{k}_3	$S_{\hat{k}_{\text{GCV}}}(h_3)$	$S(\hat{h}_{\text{AICc}})$
Gaussian	385	0.0231	27	0.0254	4	0.0368	0.04857
student	360	0.0221	25	0.0262	4	0.0353	0.05199
Function m_2							
Gaussian	330	0.0477	128	0.0581	14	0.0782	0.1175
student	1621	0.0563	160	0.0660	16	0.0754	0.1184
Function m_3							
Gaussian	30	0.0017	7	0.0016	2	0.0014	0.00178
student	29	0.0017	8	0.0016	2	0.0016	0.0018

TABLE 2

Median over 100 simulations of the number of iterations and the MSE for Gaussian kernel smoother, $n = 50$ data points. $S(\hat{h}_{\text{AICc}})$ denotes the bandwidth chosen by the modified AIC criteria.

The simulation results reported in the above tables show that the iterative bias reduction scheme works well in practice, even for moderate sample sizes. While starting with a very smooth pilot requires more iterations, the mean squared error of the resulting smoother is somewhat smaller compared to a more noisy initial smoother. Figures 5 and 6 also support this claim.

7. Discussion. In this paper, we make the connection between iterative bias correction and the L_2 boosting algorithm, thereby providing a new interpretation for the latter. A link between bias reduction and boosting was suggested by [30] in his discussion of the seminal paper [17], and explored in Di Marzio and Taylor [8, 9] for the special case of kernel smoothers. In this paper, we show that this interpretation holds for general linear smoothers.

It was surprising to us that not all smoothers were suitable to be used for boosting. We show that many weak learners, such as the k -nearest neighbor smoother and some kernel smoothers, are not stable under L_2 boosting. Our results extend and complement the recent results of Di Marzio and Taylor [9].

Iterating the boosting algorithm until convergence is not desirable. Better smoothers result if one stops the iterative scheme. We have explored, via simulations, various data driven stopping rules and have found that for the linear smoothers, the Generalized Cross Validation criteria works very well, even for moderate sample sizes of 50. In our simulations show that optimally correcting the bias (by stopping the L_2 boosting algorithm after a suitable number of iterations) produced better smoothers than the one with the best data-dependent smoothing parameter.

Finally, the iterative bias correction scheme can be readily extended to multivariate covariates X , as in Buhlmann [3].

APPENDIX A: APPENDIX

Proof of Theorem 2.1 To show 2.5, let $\Sigma = I + (I - S) + \dots + (I - S)^{k-1}$. The conclusion follows from a telescoping sum argument applied to

$$S\Sigma = \Sigma - (I - S)\Sigma = I - (I - S)^k.$$

Proof of Theorem 2.3

$$\begin{aligned} \|\hat{b}_k\|^2 &= \|(I - S)^{k-1}SY\|^2 \\ &= \|(I - S)(I - S)^{k-2}SY\|^2 \leq \|(I - S)\|^2 \|\hat{b}_{k-1}\|^2 \\ &\leq \|\hat{b}_{k-1}\|^2, \end{aligned}$$

where the last inequality follows from the assumptions on the spectrum of $I - S$. Similarly, one shows that

$$\|R_k\|^2 = \|(I - S)^kY\|^2 \leq \|I - S\|^2 \|R_{k-1}\|^2 < \|R_{k-1}\|^2.$$

Proof of Theorem 3.1 To simplify the exposition, let us assume that the X_i 's are ordered. Let us consider the K -nn smoother the matrix S is of general term

$$S_{ij} = \frac{1}{K} \quad \text{if } X_j \in K\text{-nn}(X_i).$$

In order to bound the singular values of $(I - S)$, consider the eigen values of $(I - S)(I - S)'$ which are the square of the singular values of $I - S$. Since $A = (I - S)(I - S)'$ is symmetric, we have for any vector x that

$$(A.1) \quad \lambda_n \leq \frac{x'Ax}{x'x} \leq \lambda_1.$$

Let us find a vector x such that $x'Ax > x'x$. First notice that

$$A_{ii} = 1 - \frac{1}{K}.$$

Next, consider the vector x of \mathbb{R}^n that is zero every where except at position $(i - l_1)$ (respectively i and $i + l_2$) where its value is -1 (respectively 2 and -1). For this choice, we expand $x'Ax$ to get

$$\begin{aligned} x'Ax &= A_{i-l_1, i-l_1} + 4A_{i,i} + A_{i+l_2, i+l_2} - 4A_{i-l_1, i} - 4A_{i, i+l_2} + 2A_{i+l_2, i-l_1} \\ &= 6 - \frac{6}{K} - 4A_{i-l_1, i} - 4A_{i, i+l_2} + 2A_{i+l_2, i-l_1}. \end{aligned}$$

To show that this last quantity is larger than $x^t x = 6$, we need to suitably bound the off-diagonal elements of $A = I - S - S' + SS'$. To bound A_{ij} , where $j = i + l$ and $l < K$, we need to consider three cases:

1. If X_i belongs to the K -nn of X_j and vice versa, then $S_{ij} = S'_{ji} = 1/K$. This does not mean that all the K -nn neighbor of X_i are the same as those of X_j , but if it is the case, then $(SS')_{ij} \leq K/K^2$ and otherwise in the pessimistic case, we bound $(SS')_{ij} \geq (l + 1)/K^2$. It therefore follows that

$$(l + 1)/K^2 - \frac{2}{K} \leq A_{i,i+l} \leq \frac{K}{K^2} - \frac{2}{K} = -\frac{1}{K}.$$

2. If X_i belongs to the K -nn of X_j $S_{ij} = 1/K$ but X_j does not belong to the K -nn of X_i then $S'_{ji} = 0$. There is at a maximum of $K - 1$ points that are in the K -nn of X_i and in the K -nn of X_j so $(SS')_{ij} \leq (K - 1)/K^2$. In the pessimistic case, there is only one point, which leads to the bound

$$\frac{1}{K^2} - \frac{1}{K} \leq A_{i,i+l} \leq \frac{K - 1}{K^2} - \frac{1}{K} \leq -\frac{1}{K^2}.$$

3. If X_i does not belong to the K -nn of X_j $S_{ij} = 0$ and X_j does not belong to the K -nn of X_i then $S'_{ji} = 0$. However there are potentially as many as $l - 1$ points that are in the K -nn of X_i and in the K -nn of X_j , so that $(SS')_{ij} \leq (l - 1)/K^2$. In that case

$$0 \leq A_{ij} \leq \frac{l - 1}{K^2} \leq \frac{K - 2}{K^2}.$$

With these bounds for the off-diagonal terms, we are able to major $x'Ax$.

Before continuing, we need to discuss the relative position of the points X_{i-l_1} , X_i and X_{i+l_2} . We chose them such that

$$X_{i-l_1} \in K\text{-nn}(X_i) \quad \text{and} \quad X_i \in K\text{-nn}(X_{i-l_1}).$$

For this choice, we calculate

$$\begin{aligned} \frac{l_1 + 1 - 2K}{K^2} &\leq A_{i-l_1,i} \leq -\frac{1}{K} \\ \frac{l_2 + 1 - 2K}{K^2} &\leq A_{i,i+l_2} \leq -\frac{1}{K}, \end{aligned}$$

so that

$$6 - \frac{6}{K} + \frac{8}{K} + 2A_{i+l,i-l} \leq x'Ax \leq 6 + \frac{2}{K} + 2A_{i+l,i-l}.$$

The latter shows that $x'Ax > x'x$ whenever

$$A_{i+l_2, i-l_1} > -\frac{1}{K},$$

which is always true if the condition

$$X_{i-l_1} \notin K\text{-nn}(X_{i+l_2}) \quad \text{or} \quad X_{i+l_2} \notin K\text{-nn}(X_{i-l_1})$$

is satisfied because in such case, we have

$$-\frac{1}{K} < A_{i-l_1, i+l_2} \leq \frac{1}{K^2}.$$

Proof of Theorem 3.2 Let X_1, \dots, X_n is an i.i.d. sample from a density f that is bounded away from zero on a compact set strictly included in the support of f . Consider without loss of generality that $f(x) \geq c > 0$ for all $|x| < b$.

We are interested in the sign of the quadratic form $u^t Au$ where the individual entries A_{ij} of matrix A are equal to

$$A_{ij} = \frac{K_h(X_i - X_j)}{\sqrt{\sum_l K_h(X_i - X_l)} \sqrt{\sum_l K_h(X_j - X_l)}}.$$

Recall the definition of the scaled kernel $K_h(\cdot) = K(\cdot/h)/h$. If v is the vector of coordinate $v_i = u_i / \sqrt{\sum_l K_h(X_i - X_l)}$ then we have $u^t Au = v^t \mathbb{K} v$, where \mathbb{K} is the matrix with individual entries $K_h(X_i - X_j)$. Thus any conclusion on the quadratic form $v^t \mathbb{K} v$ carry on to the quadratic form $u^t Au$.

To show the existence of a negative eigenvalue for \mathbb{K} , we seek to construct a vector $U = (U_1(X_1), \dots, U_n(X_n))$ for which we can show that the quadratic form

$$U^t \mathbb{K} U = \sum_{j=1}^n \sum_{k=1}^n U_j(X_j) U_k(X_k) K_h(X_j - X_k)$$

converges in probability to a negative quantity as the sample size grows to infinity. We show the latter by evaluating the expectation of the quadratic form and applying the weak law of large number.

Let $\varphi(x)$ be a real function in L_2 , define its Fourier transform

$$\hat{\varphi}(t) = \int e^{-2i\pi tx} \varphi(x) dx$$

and its Fourier inverse by

$$\hat{\varphi}_{inv}(t) = \int e^{2i\pi tx} \varphi(x) dx.$$

For kernels $K(\cdot)$ that are real symmetric probability densities, we have

$$\hat{K}(t) = \hat{K}_{inv}(t).$$

From Bochner's theorem, we know that if the kernel $K(\cdot)$ is not positive definite, then there exists a bounded symmetric set A of positive Lebesgue measure (denoted by $|A|$), such that

$$(A.2) \quad \hat{K}(t) < 0 \quad \forall t \in A.$$

Let $\hat{\varphi}(t) \in L_2$ be a real symmetric function supported on A , bounded by B (i.e. $|\hat{\varphi}(t)| \leq B$). Obviously, its inverse Fourier transform

$$\varphi(x) = \int_{-\infty}^{\infty} e^{-2\pi i x t} \hat{\varphi}(t) dt$$

is integrable and by virtue of Parseval's identity

$$\|\varphi\|^2 = \|\hat{\varphi}\|^2 \leq B^2 |A| < \infty.$$

Using the following version of Parseval's identity [see 14, p.620]

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \varphi(x) \varphi(y) K(x-y) dx dy = \int_{-\infty}^{\infty} |\hat{\varphi}(t)|^2 \hat{K}(t) dt,$$

which when combined with equation (A.2), leads us to conclude that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \varphi(x) \varphi(y) K(x-y) dx dy < 0.$$

Consider the following vector

$$U = \frac{1}{nh} \begin{bmatrix} \frac{\varphi(X_1/h)}{f(X_1)} \mathbb{I}(|X_1| < b) \\ \frac{\varphi(X_2/h)}{f(X_2)} \mathbb{I}(|X_2| < b) \\ \vdots \\ \frac{\varphi(X_n/h)}{f(X_n)} \mathbb{I}(|X_n| < b) \end{bmatrix}.$$

With this choice, the expected value of the quadratic form is

$$\begin{aligned} \mathbb{E}[Q] &= \mathbb{E} \left[\sum_{j,k=1}^n U_j(X_j) U_k(X_k) K_h(X_j - X_k) \right] \\ &= \frac{1}{n} \int_{-b}^b \frac{1}{f(s)h^2} \varphi(s/h)^2 K_h(0) ds \\ &\quad + \frac{n^2 - n}{n^2} \int_{-b}^b \int_{-b}^b \frac{1}{h^2} \varphi(s/h) \varphi(t/h) K_h(s-t) ds dt \\ &= I_1 + I_2. \end{aligned}$$

We bound the first integral

$$\begin{aligned}
I_1 &= \frac{K_h(0)}{nh^2} \int_{-b}^b \frac{\varphi(s/h)^2}{f(s)} ds \\
&\leq \frac{K_h(0)}{nch} \int_{-b/h}^{b/h} \varphi(u)^2 du \\
&\leq \frac{B^2|A|K(0)}{ch^2} n^{-1}.
\end{aligned}$$

Observe that for any fixed value h , the latter can be made arbitrarily small by choosing n large enough. We evaluate the second integral by noting that

$$\begin{aligned}
I_2 &= \left(1 - \frac{1}{n}\right) h^{-2} \int_{-b}^b \int_{-b}^b \varphi(s/h) \varphi(t/h) K_h(s-t) ds dt \\
&= \left(1 - \frac{1}{n}\right) h^{-2} \int_{-b}^b \int_{-b}^b \varphi(s/h) \varphi(t/h) \frac{1}{h} K\left(\frac{s}{h} - \frac{t}{h}\right) ds dt \\
(A.3) \quad &= \left(1 - \frac{1}{n}\right) h^{-1} \int_{-b/h}^{b/h} \int_{-b/h}^{b/h} \varphi(u) \varphi(v) K(u-v) du dv.
\end{aligned}$$

By virtue of the dominated convergence theorem, the value of the last integral converges to $\int_{-\infty}^{\infty} |\hat{\varphi}(t)|^2 \hat{K}(t) dt < 0$ as h goes to zero. Thus for h small enough, (A.3) is less than zero, and it follows that we can make $\mathbb{E}[Q] < 0$ by taking $n \geq n_0$, for some large n_0 . Finally, convergence in probability of the quadratic form to its expectation is guaranteed by the weak law of large numbers for U statistics (see Grams and Serfling [18] for example). The conclusion of the theorem follows.

Proof of Proposition 3.3 We are interested in the sign of the quadratic form $u^t \mathbb{K} u$ (see proof of Theorem 3.2). Recall that if \mathbb{K} is semidefinite then all its principal minor [see 23, p.398] are nonnegative. In particular, we can show that A is non-positive definite by producing a 3×3 principal minor with negative determinant. To this end, take the principal minor $\mathbb{K}[3]$ obtained by taking the rows and columns (i_1, i_2, i_3) . Without loss of generality, let us assume that $X_{i_1} < X_{i_2} < X_{i_3}$. The determinant of $\mathbb{K}[3]$ is

$$\begin{aligned}
\det(\mathbb{K}[3]) &= K_h(0) \left[K_h(0)^2 - K_h(X_{i_3} - X_{i_2})^2 \right] \\
&\quad - K_h(X_{i_2} - X_{i_1}) \\
&\quad \times [K_h(0)K_h(X_{i_2} - X_{i_1}) - K_h(X_{i_3} - X_{i_2})K_h(X_{i_3} - X_{i_1})] \\
&\quad + K_h(X_{i_3} - X_{i_1}) \\
&\quad \times [K_h(X_{i_2} - X_{i_1})K_h(X_{i_3} - X_{i_2}) - K_h(0)K_h(X_{i_3} - X_{i_1})].
\end{aligned}$$

Let us evaluate this quantity for the uniform and Epanechnikov kernels.

Uniform kernel. Let h be larger than the minimum distance between three consecutive points, and chose the index i_1, i_2, i_3 such that

$$X_{i_2} - X_{i_1} < h, \quad X_{i_3} - X_{i_2} < h, \quad \text{and} \quad X_{i_3} - X_{i_1} > h.$$

With this choice, we readily calculate

$$\det(\mathbb{K}[3]) = 0 - K_h(0) [K_h(0)^2 - 0] - 0 < 0.$$

Since a principal minor of \mathbb{K} is negative, we conclude that \mathbb{K} and A are not semidefinite positive.

Epanechnikov kernel. For i_1, i_2, i_3 fixed, denote by $x = X_{i_2} - X_{i_1}$ and by $y = X_{i_3} - X_{i_2}$, and assume that $h > \min(x, y)$. The determinant $\det(\mathbb{K}[3])$ is a bivariate function of x and y (as $X_{i_3} - X_{i_1} = x + y$). Numerical evaluations of that function show that as soon as we have the range of the three points less than the bandwidth, the determinant of $\mathbb{K}[3]$ is negative.

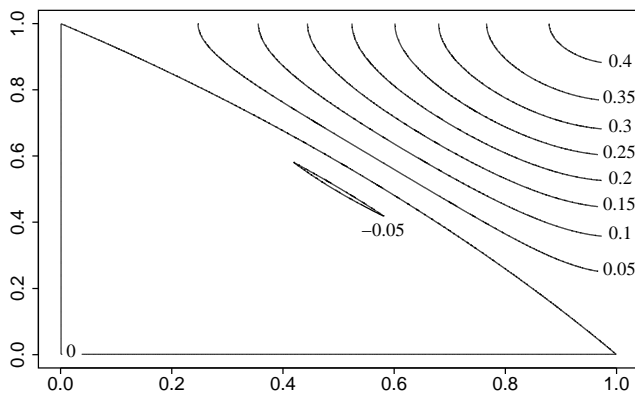


FIG 9. Contour of $\det(\mathbb{K}[3])$ as a function of (x, y) .

Thus a principal minor of \mathbb{K} is negative, and as a result, \mathbb{K} and A are not semidefinite positive.

REFERENCES

- [1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and B. F. Csaki, editors, *Second international symposium on information theory*, pages 267–281, Budapest, 1973. Akademiai Kiado.
- [2] L. Breiman. Arcing classifier (with discussion). *Ann. of Statist.*, 26:801–849, 1998.
- [3] P. Bühlmann. Boosting for high-dimensional linear models. *Ann. of Statist.*, 34: 559–583, 2006.

- [4] P. Bühlmann and B. Yu. Boosting with the l_2 loss: Regression and classification. *J. Amer. Statist. Assoc.*, 98:324–339, 2003.
- [5] A. Buja, T. Hastie, and R. Tibshirani. Linear smoothers and additive models. *Ann. of Statist.*, 17:453–510, 1989.
- [6] W. Cleveland. Robust locally weighted regression and smoothing scatterplots. *J. Amer. Stat. Ass.*, 74:829–836, 1979.
- [7] P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403, 1979.
- [8] M. Di Marzio and C. Taylor. Boosting kernel density estimates: a bias reduction technique ? *Biometrika*, 91:226–233, 2004.
- [9] M. Di Marzio and C. Taylor. Multiple kernel regression smoothing by boosting. *submitted*, 2007.
- [10] M. Di Marzio and C. Taylor. On boosting kernel regression. *to appear in JSPI*, 2008.
- [11] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression (with discussion). *Ann. of Statist.*, 32:407–451, 2004.
- [12] R. Eubank. *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New-York, 1988.
- [13] J. Fan and I. Gijbels. *Local Polynomial Modeling and Its Application, Theory and Methodologies*. Chapman et Hall, New York, 1996.
- [14] W. Feller. *An introduction to probability and its applications*, volume 2. Wiley, New York, 1966.
- [15] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121:256–285, 1995.
- [16] J. Friedman. Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, 28(337-407), 2001.
- [17] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Ann. of Statist.*, 28:337–407, 2000.
- [18] W. Grams and R. Serfling. Convergence rates for u-statistics and related statistics. *Annals of Statistics*, 1:153–160, 1973.
- [19] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1995.
- [20] N. Hengartner and E. Matzner-Løber. Asymptotic unbiased density estimators. *accepted in ESAIM*, 2007.
- [21] N. Hengartner, M. Wegkamp, and E. Matzner-Løber. Bandwidth selection for local linear regression smoothers. *JRSS B*, 64:1–14, 2002.
- [22] N. Hjort and I. Glad. Nonparametric density estimation with a parametric start. *Ann. Statist.*, 23:882–904, 1995.
- [23] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge, New York, 1985.
- [24] C. Hurvich, G. Simonoff, and C. L. Tsai. Smoothing parameter selection in nonparametric regression using and improved akaike information criterion. *J. R. Statist. Soc. B*, 60:271–294, 1998.
- [25] M. Jones, O. Linton, and J. Nielsen. A simple and effective bias reduction method for kernel density estimation. *Biometrika*, 82:327–338, 1995.
- [26] K. Li. From stein’s unbiased risk estimate to the method of generalized cross validation. *Ann. Statist.*, 13:1352–1377, 1985.
- [27] K.-C. Li. Asymptotic optimality for C_p , C_L , cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, 15:958–975, 1987.
- [28] C. L. Mallows. Some comments on C_p . *Technometrics*, 15:661–675, 1973.
- [29] E. Nadaraya. On estimating regression. *Theory Probability and their Applications*, 9 (134-137), 1964.
- [30] G. Ridgeway. Additive logistic regression: a statistical view of boosting: Discussion.

- Ann. of Statist.*, 28:393–400, 2000.
- [31] R. Schapire. The strength of weak learnability. *Machine learning*, 5:197–227, 1990.
 - [32] L. Schwartz. *Analyse IV applications à la théorie de la mesure*. Hermann, Paris, 1993.
 - [33] G. Schwarz. Estimating the dimension of a model. *Annals of statistics*, 6:461–464, 1978.
 - [34] J. Simonoff. *Smoothing Methods in Statistics*. Springer, New York, 1996.
 - [35] J. Tukey. *Explanatory Data Analysis*. Addison-Wesley, 1977.
 - [36] F. Utreras. Natural spline functions, their associated eigenvalue problem. *Numer. Math.*, pages 107–117, 1983.
 - [37] G. Wahba. *Spline models for observational data*. SIAM, Philadelphia, 1990.
 - [38] G. Watson. Smooth regression analysis. *Sankhya A*, 26(359-372), 1964.
 - [39] E. Whittaker. On a new method of graduation. *Proc. Edinburgh Math; Soc.*, 41: 63–75, 1923.

ADDRESS OF P-A CORNILLON
 UMR ASB - MONTPELLIER SUPAGRO
 34060 MONTPELLIER CEDEX 1
 E-MAIL: pierre-andre.cornillon@supagro.inra.fr

ADDRESS OF N. HENGARTNER
 LOS ALAMOS NATIONAL LABORATORY,
 NW, USA
 E-MAIL: nickh@lanl.gov

ADDRESS OF E. MATZNER-LØBER
 STATISTICS, IRMAR UMR 6625,
 UNIV. RENNES 2,
 35043 RENNES, FRANCE
 E-MAIL: eml@uhb.fr